Univerzita Karlova v Praze

Filozofická fakulta

Ústav anglického jazyka a didaktiky

Název práce: Klíčová slova v románech J. K. Rowlingové o Harry Potterovi

Title of Thesis: Keywords in J. K. Rowling's Harry Potter novels

BAKALÁŘSKÁ PRÁCE

Zpracovala: Karolina Hošek

Vedoucí práce: PhDr. Markéta Malá, Ph.D.

Srpen 2011

„Prohlašuji, že jsem bakalářskou práci vypracovala samostatně a že jsem uvedla všechny použité prameny a literaturu. Souhlasím se zapůjčením bakalářské práce ke studijním účelům.“

“I declare that the following BA thesis is my own work for which I used only the sources and literature mentioned. I have no objections to the BA thesis being borrowed and used for study purposes.”

V Praze dne 13.08.2011,

Karolina Hošek

Acknowledgment

Tímto bych chtěla poděkovat svojí školitelce, PhDr. Markétě Malé Ph.D., za její trpělivé vedení, nedocenitelné rady a smysl pro humor. Dále bych chtěla poděkovat všem zástupcům Ústavu anglického jazyka a didaktiky, jejichž výuka mi v průběhu studia poskytla řadu znalostí, bez kterých by tato práce zcela jistě nevznikla. V neposlední řadě bych také chtěla poděkovat Bc. Markovi Leškovi, jehož výzkumná činnost v oblasti lingvistiky mi byla velkou inspirací, a rodině a přátelům, kteří mi v době zpracovávání práce byli oporou.

Abstract

Keyword analysis is a method used in corpus linguistics, in which keywords are generated automatically from the text against the background of a reference corpus serve as the starting point of the analysis of the text. In this thesis, keywords are retrieved from a literary text and analyzed, with a focus on their function as style markers. The theoretical part of this work centers around the description of corpus stylistics, the role of keywords in corpus stylistics and previous findings in the field by other researchers, including linguistic analyses of literary texts. This chapter is complemented by a methodological part which describes the software used in the research and its individual tools.

The keyword analysis method is tested by comparing the seven novels of the Harry Potter series to a reference corpus – a subcorpus of the British National Corpus comprising fiction for children and teenage readers. The first one hundred keywords generated using the concordance software AntConc are further divided into subgroups of grammatical and stylistic words, and lexical words, which are listed together with proper nouns. The analyzed keywords illustrate the ways in which the explored text differs from comparable works of children's literature in both lexis and grammar. The concluding part evaluates the method and its use in the exploration of literary works based on the practical part of this study.

Analýza klíčových slov je výzkumnou metodou korpusové lingvistiky, při které jsou z výchozího textu specializovaným software automaticky generována klíčová slova oproti srovnávacímu referenčnímu korpusu, která slouží jako výchozí bod pro další analýzu textu. Tato bakalářská práce se zabývá popisem klíčových slov jakožto stylisticky významných kompoment textu a jejich následným generováním z literárního textu a podrobnou klasifikací. Teoretická část práce se dále soustředí na obor korpusové stylistiky, význam klíčových slov pro tuto disciplínu a shrnutí dosavadních projektů v této oblasti, včetně lingvistických rozborů literárních textů. Tento oddíl je doplněný metodologickou částí, která popisuje užitý software a možnosti jeho využití.

Metoda je ověřována na korpusu sestávajícím se ze sedmi dílů série knih o Harrym Potterovi, který je porovnávan s referenčním korpusem – subkorpusem Britského národního korpusu složeným z textů pro děti a mládež. Prvních sto klíčových slov získaných pomocí konkordančního programu AntConc je dále rozděleno na slova gramatická a stylistická a slova lexikální, kterým jsou přiřazena i vlastní jména. Analyzovaná klíčová slova dobře ukazují, čím se zkoumaný text liší od podobně orientované dětské literatury v oblasti lexika a gramatických konstrukcí. Na základě této praktické studie jsou následně zhodnoceny možnosti metody a jejího užití pro zkoumání literárních textů.

Contents

Tables and examples

Abbreviations

Anthony – Laurence Anthony, the author of the AntConc software and manual
BNC – British National Corpus
CGEL – A Comprehensive Grammar of English Language
Potter – corpus composed of the seven novels of the Harry Potter series by J.K.Rowling
Teen-BNC – subcorpus of BNC comprising texts of literature for children and teenagers published between 1960 and 1993

1. Introduction

Keyword analysis is a relatively new research method, though it has been steadily gaining importance over the past two decades due to rapid development of the computational software used in corpus linguistics. Unlike basic frequency lists, the keyword listing allows to detect statistically relevant items which appear more prominent in comparison with the reference corpus, therefore allowing deeper insight into the text and identification of topics or data that would be potentially overlooked by traditional literary critical analysis. The first part of this paper addresses the issues related to the definition of keywords and the history of keyword analysis in the stylistic study of text; it is complemented by a methodological section concerned especially with the compilation of the reference corpus, the choice of software for generating the keywords and problems that might be encountered in the process. The second part of the thesis addresses the data retrieved in the research and attempts to answer the question of their relevance for stylistic analysis and potential benefits of this technique. Since the macroscopic keyword analysis method is strictly based on the empirical approach to data, it can reveal the grammatical patterns and structural features not so apparent in intuitive reading that relies on the identification of the prominent topics through the plot and dominant stock of lexical words of the text. The keyword analysis may be therefore very beneficial in the process of decoding the novel's literary meaning. This paper aims to explore the possibilities of using the quantitative method of keyword analysis as grounds to further stylistic analysis of the text and the role of keywords in such, showing how certain lexical features or groups of features characterize a text.

2. Keywords

Keywords can be defined as "words which occur with significantly greater frequency in one part of the corpus than another." (Hoffmann, 2008: 203) There is, however, no clear-cut and universally applicable definition of keywords because each study adjusts the concept to meet the purposes of the paper.

2.1 What are keywords

A plausible definition of keyness as an attribute ultimately possessed by all keywords is 'a quality words may have in given text or set of texts, suggesting that they are important, they reflect what the text is really about, avoiding trivia and insignificant detail.' (Scott & Tribble, 2006: 55-56) Furthermore, Scott also employs the notion of key-key words which are the lexical items that are key in all or in a significant amount of texts in the corpus that is being investigated. (Rayson, 2008: 523) It is "a matter of being statistically unusual relative to some norm." (Culpeper, 2009: 34) Culpeper links the concept to that of distributional style markers, saying that "words whose frequencies differ significantly from their frequencies in a norm are precisely what keywords are." (Culpeper, 2009: 33) Similarly to this assumption, Fischer-Starcke sees keywords as the indicators of the "dominant topics or themes of a text or a corpus since

the reason for their frequent occurrence in the data is their significance either for the data's content or its structure." (Fischer-Starcke, 2009: 496)

2.2 What are not keywords

The keywords treated in this paper are always understood in the sense of textual keywords, not cultural keywords in the sense of "sociologically important words, what one might call focal or pivotal words" (Firth, 1935 cited in Bondi 2010: 2) neither those identified by intuition which are labeled as 'key' "because they are of particular social, (…) or political significance." (Culpeper, 2009: 32) They are certainly not to be understood as they are frequently used in the knowledge management, the words that "help identify a text in structured databases, such as for example library resources." (Bondi, 2010: 5) Similarly, they are not to be confused with the term *keyword* as used in colloquial speech, denoting simply something of importance perhaps completely out of any textual context.

2.3 Identification of keywords

The identification of keywords is based on frequency of occurrence. It takes a frequency-ordered word-list for the study corpus as its starting point. "Simple verbatim lexical repetition alone will not do, however" (Scott & Tribble, 2006: 58) as the top frequency words are typically grammatical words (such as determiners and prepositions) and high-frequency lexical items (such as *time*, *know*, *people*) which can hardly serve as indicators of 'aboutness'. The calculation of keywords therefore "requires a 'reference corpus word-list' which can indicate how often any given word can be expected to occur in the language or genre in question. This will be used as a filter." (ibid)

2.3.1 The choice of the reference corpus

In any keyword analysis research, it is necessary to create a set of data for comparison – the reference corpus. Because keyness is, in theory, simply a matter of contrast, such corpus is then used as a filter serving to separate the items of interest. There is, however, "no standard recipe for the composition of a special corpus." (Teubert&Čermáková, 2007: 69) In each individual study it is necessary to make various adjustments for the corpus to fit the research focus.

2.3.2 Calculation of keyness

The words are considerd key if the difference in their textual frequency in the study corpus and the reference corpus is statistically significant – "in other words, if we can say with a sufficient degree of confidence that the observed difference is not due to chance" (Hoffmann et al., 2008: 204) The statistical measure used to compare the two word-lists is typically either a chi-square test or log-likelihood test. These two methods are related, and therefore share the same critical values (cf. Table 1 below).

Table 1: Keyness values and levels of significance

critical value	implies the difference is significant at the	
3.841	5 % level	$p < 0.05$
6.635	1 % level	$p < 0.01$
10.827	0.1 % level[1]	$p < 0.001$

2.3.3 Possible problems and drawbacks

While compiling the reference corpus, it is rather difficult to establish the adequate size of the corpus and source of the input data. In order to produce relevant keyword lists, the two sets of data have to share certain common grounds. Fischer Starcke (2009) suggests that in order to avoid bias in the words generated as key the reference corpus should ideally match the explored text in content and be also of adequate size. She also recommends comparing two or more lists of keywords retrieved by using different reference corpora to identify a potential bias and avoid it, a method that has proved itself efficient in her research. The final choice of the reference corpus data and their size will have, however, a direct influence on what keyword lists are generated. Large general corpora are often used as reference corpora, e.g. Stubbs (2005) compares Conrad's *Heart of Darkness* with the one-million written component of the British National Corpus Sampler. There is, nonetheless, a certain advantage to having a small/medium sized data set because it allows careful scrutiny and manual checking of the results and investigation of the concordance lines, and the distinguishing stylistic features of the literary text are not 'swallowed'by the enormous sets of data of the large corpora.

2.4 Types of keywords

Once we have identified the keywords, we may further divide them into several subtypes. The most fundamental distinction can be made on the basis of their function within the text, splitting them into two uneven groups (cf. Scott, 2000, 2008). Bondi (2010) compares this division to Sinclair and Mauranen's distinction between "message oriented elements" in text directed towards topical development and "organization-oriented elements" which make the content structured. (Sinclair&Mauranen, 2006 cited in Bondi 2010: 7) Some researchers recognize a third group of keywords that have "little semantic content, but rather pragmatic import and they tend to be peripheral to syntax," (Culpeper, 2009: 39) therefore not falling into either of the two categories. Halliday (1973, 1978, 1994) performs further division of keywords, adding the category of "interpersonal keywords". Keywords may be also further classified based on the fact

[1] The significance at the 0.1% level means "that there is only a 0.1% probability that the difference between the two subcorpora is due to chance alone – thus suggesting that the observed differences have to do with the nature of the two subcorpora rather than with random variation." (Hoffmann et al., 2008: 208)

whether they are positive or negative, i.e. unusually frequent or unusually infrequent. This distinction and its importance will be dealt with in more detail in the methodological part of the text as it is closely connected with selecting the working settings of the software. For the purposes of this paper, a two-way distinction will be adopted with the "interpersonal keywords" listed along with grammatical words.

2.4.1 Lexical keywords

The largest stock of lexical units as to the relative amount and diversity they present usually belongs to the group of lexical words, these are mostly open-class words and the carriers of stress and content. They are indicators of the text's "aboutness"(cf. Culpeper 2009 or Scott 2000, 2008). It is very likely that the prominent items in this category will be nouns denoting some very general categories such as *people*, *place*, *time, issue* (cf. Scott and Tribble, 2006) that are often means of substitution of more concrete terms but standing on their own do not convey too much of a information; they are important means of lexical cohesion but do not reflect the important themes of the text. The words that are of the category of lexical keywords will be more likely genre-specific content words, very concrete in their denotational meaning.

2.4.2 Proper nouns

For the purpose of this study, the proper nouns will be listed along with lexical keywords even though they may carry some specific features which distinguish them from other lexical words. These items often have very high value of keyness due to only limited cohesive function of pronominalization over longer stretches of text – repetition is eventually needed. Whereas Scott suggest that they are largely unimportant, since 'a text about racing could wrongly identify as key, names of the horses which are quite incidental to the story' (Scott, 2008 cited in Culpeper 2009: 38). Culpeper claims that 'in fictional text, they may be of certain interest, as they relate to the key aspects of the fictional world (Culpeper, 2009: 38). This is certainly true, but they also frequently only stress what is noticeable by intuitive observation e.g. negative frequency of a character's name when comparing to sequels of one novel may signify that he or she died (or simply came out of center of attention). As observed by Culpeper, a proper name as a negative keyword in the identical character's speech may be a reasonable clue to assume that he or she is a frequent address or reference for the other characters. (ibid) These are however not the main focus of this paper and will be therefore discussed only when relevant in the larger context of stylistic analysis.

2.4.3 Grammatical and stylistic keywords

The other group of words which are "indicators more of style than of 'aboutness'" (Scott, 2008 cited in Culpeper, 2009: 38) may as well fall under the umbrella term 'grammatical'. Culpeper suggests that this assortment of words includes all the items which do not belong to the first category and are significant namely for the organization of the text, providing textual cohesion by linking the elements of the surface structure of

the text, putting the lexical units of the former group in relation to each other and also placing the text itself in relation to its author or receiver. These are very frequently but not always unstressed, closed-class words. Bondi (2010) stresses their importance as they are "signals of organizational structure [and] will thus be key (or "pivotal") in reading because they facilitate access to the information required." Despite their primary structural function, they are not always completely devoid of meaning: In his study on Hamlet character's talk, Culpeper (2009) suggests that the frequent use of the keyword *if* may be, for example, directly linked to the character's unconcealed anxiety and hesitation and the recurrent use of the subjunctive mood similarly serves to convey emotional utterances expressing a wish, a regret etc.

3. Applying keyword analysis to works of literature

Prior to analyzing the benefits and potential limitations inevitably connected with this research method, it is necessary to clarify the approach taken in this study. Unlike corpus-based analyses that center around a previously formulated hypothesis that is consequently verified using corpora which is only "brought in as an extra bonus rather than as determining factor" and "is never in the position to challenge" (Tognini-Bonelli, 2001: 66) the pre-established categories as stated by the linguist, this work takes a corpus-driven approach. This means that no hypotheses apart from very general statements about the research methods and previous findings in the field are formed prior examining the text itself. This is to say that it is the results delivered by the corpora that are interpreted and "recurrent patterns and frequency distributions are expected to form the basic evidence for linguistic categories; the absence of a pattern is considered potentially meaningful." (Tognini-Bonelli, 2001: 84) While discussing the advantages and disadvantage of using corpora in linguistic exploration of a literary text, most statements that will be made should, nevertheless, apply to both approaches that may be taken in such study.

3.1 The benefits of linguistic analyses of literary texts

The corpus driven and corpus based linguistic and stylistic analysis of the literary work has numerous advantages. It is, at least in its opening phase, a very speedy process thanks to the development of advanced computational software which makes the task both easy and efficient. A variety of additional features such as tagging of semantic domains or word classes allows organizing and listing the retrieved word stock according to one's research purpose. For the first time in history, it was now made possible to take a quantitative and empirical approach to larger stretches of text which were before neglected simply because of the amount of the data to be dealt with. Such computerized analysis is capable of grasping the high-frequency items among grammatical words and borderline discourse markers which are unlike their lexical counterparts almost imperceptible by intuitive human interpretation of a text. Elucidating the structural features which organize and hold together the text along with the content words is an excellent starting point for further stylistic analysis of such. The results of such analysis may be employed as a useful tool in translation as was

suggested for example by Čermáková and Fárová (2010). It may as well serve for the authorship attribution as described in Archer (2009). In her study on practical use of corpora in the linguistic research, Bonelli (2001) highlights several advantages this research method brings about, such as allowing "parole becoming amenable to systematic observation" and paving the way for "a qualitative change in the description of language." (Bonelli, 2001: 86) Mahlberg (2009) also points out the obvious advantage of corpus stylistics: she understands the use of the method "as a way to add systemticity to stylistic analysis," and adds that it may as well serve in reducing subjectivity, inevitably present in intuitively performed analysis. (Mahlberg, 2009: 48) She concludes her observation saying that the methodology can also enrich stylistics by adding "categories developed in the field of corpus linguistics (e.g., semantic prosodies) to the inventory of stylistic description." (Mahlberg, 2009: 50)

3.2 Problems and limitations of linguistic analyses of literary texts

There are undoubtedly certain constraints to taking strictly empirical approach to the collected data; these limitations have been pointed out in previous research. Apart from rather technical limitations, it has been acknowledged that despite the rapid progress in the field a researcher more then often enters unexplored grounds (cf. Tognini-Bonelli, 2001) not so much when it comes to the retrieval but more often in an attempt to interpret the data. In spite of being praised for reducing subjectivity in the exploration (cf. Mahlberg, 2009 or Fischer-Starcke 2009), the resultant data always require non-computerized interpretation by a researcher so that the final and perhaps most important part of the process is the most intuitive and therefore the least objective one. Culpeper warns against performing 'a keyword analysis in a relatively mechanical way without a critical awareness of what is being revealed of how it is being revealed.'(Culpeper, 2009: 30) Similarly, Tognini-Bonelli (2001) advises that such research process, however computerized, needs to be "constantly mediated by the linguist." (Tognini-Bonelli, 2001: 85) Beber Sardinha (1999) criticizes keyword analyses for delivering more keywords than the researcher may process within a realistic time frame. (Beber Sardinha, 1999 cited in Culpeper, 2009: 31) Baker (2004) points out that 'a keyword analysis will focus only on lexical differences, not lexical similarities.'' (Baker, 2004 cited in Rayson 2008: 526) While touching upon the problems encountered investigating keyness, Scott (2010) points out the variety of statistical issues which may significantly deform the results, and recommends that the researchers proceed carefully and "take the output with discernment and discretion." (Scott, 2010: 50) In their *A Short Introduction to Corpus Linguistics* (2007), Teubert and Čermáková point out a further weakness of corpus-based and corpus-driven approaches that lies in, they claim, the fact that it "will only tell us what people have said so far. It will not tell us what people are going to say tomorrow." (Teubert&Čermáková, 2007: 47) The authors also comment on the fact that such studies depend heavily on the researcher and his "capacities to decode the linguistics patterns and grasp upon certain cues, and also on his subject of focus, his findings are somewhat limited to what he has been searching for, some significant data may be therefore omitted. (Style in Fiction, 1981: 70)

Analysing Conrad's *Heart of Darkness,* Stubbs also poits out the different roles of nouns and verbs in keyword anlaysis: "Frequent nouns may indicate superficial topics in a text …, but not its underlying themes. … Verbs are often a better candidate for stylistically relevant words." (Stubbs, 2005: 10)

3.3 Overview of theoretical approaches to the linguistic analysis of literary texts

Poetry, unlike prose, has been traditionally described from more of a linguistic point of view because of the widespread believe that „the aesthetic effect cannost be separated from the creative manipulation of the linguistic code." (Leech and Short, 1981: 2) The authors of *Style in Fiction* suggest that this may be partly caused by the fact that „the disinguishing features of a prose style tend to become detectable over longer stretches of text, and to be demonstrable ultimately only in quantitative terms" (Leech and Short, 1981: 3), therefore they are not fit for immediate linguistic analysis simply for the volume of text to be examined. There has also been a general tendency of linguists to move beyond the scope of the sentence (Leech and Short, 1981) and direct their attention on the textual and functional level of language in order to „explore the pattern and system below the surface forms of language" (Leech and Short, 1981: 5), rather than examining morphological or syntactic structures. As for stylistics, the focus is not so much on the system of rules that allow encoding the message (langue) but much more on a particular use of this system (parole). (Leech and Short, 1981: 10-11) The ‚genuine thumprint' (Leech and Short, 1981: 14) of an artist dwells below the layers of artifically stylized and manipulated text of prelearned schemes and intetionally selected lexis. There are two principal radically different views on how to look at the text and its relation to the meaning: The dualist view understands it as ‚a dress of a thought'(Leech and Short, 1981: 15), therefore admitting that there are multiple ways of conveying one message while the content remains identical. The monist view, on the contrary, is that the form inevitably influences the message, suggesting that any change in the form will ultimately (if only slightly) influence the final message. (Leech and Short, 1981: 20) This is to say, that the pervasive tendencies appearing in the text and the choices the author has made both intentionally and unintentionally will have direct effect on the outcome of his literary work. A linguistic exploration of such may, therefore, confidetly contribute to literary analyses.

3.4 Corpora and stylistics

Corpus stylistics is a relatively new discipline where a lot is to be yet discovered (cf. Baker, 2009). Corpus stylistics tries to identify and put in practice the ways in which an empirical linguistic analysis can be beneficial to the exploration of literary texts. It also attempts to establish ways in which patterns in which the writer exhibits his or her creativity and individual use of language can be observed and consequently interpreted from the linguistic point of view. The most important component of corpus stylistics is perhaps "not only the application of quantitative methods to literary texts but also a reflection on the types of questions we can ask and attempt to answer." (Mahlberg,

2009: 48) It is something of a cross-discipline whose aim is "to relate the critic's concern of aesthetic appreciation with the linguist's concern of linguistic description." (Leech and Short, 1981 cited in Mahlberg 2009, 48)

3.4.1 The role of keywords in linguistic exploration of literary works

Easily and relatively rapidly obtainable, frequency lists and keywords are very efficient as a starting point of corpus stylistic analysis. A more thorough analysis of a text may nevertheless often require employing other methods on top of the former two. The reading and detailed study of the individual concordance lines supports the interpretation and allows both verification of the data retrieved from the corpora and establishing the function or meaning of the generated keywords in context. Identifying prominent lexical item's frequent collocation which form an important part of the extended meaning of the word may also help to draw a more complex picture of what is being said. Moving beyond the scope of a single word, a researcher may turn to exploration of larger units such a lexical clusters and n-grams. All of these methods can significantly contribute to a systematic corpus stylistic analysis of a literary text.

3.5 Recent history of keyword analyses of fiction

Keyword analysis was previously applied to both non-fiction and fictional texts, with the former being the center of attention. The recent years have, however, shown a new trend – an increasing focus on corpus-driven stylistic analysis of literary texts. Keyword analysis was used by, for example, Burrows (1987) who explores the vocabularly of the characters in Jane Austen's novels, Trible (2000) studying Romantic fiction, Culpeper (2009) analyzing character-talk of Romeo and Juliet, or Fischer-Starcke (2009), who explored Austen's Pride and Prejudice and others.

3.6 Linguistic studies of the Harry Potter series

The Harry Potter series is very attractive for corpus-driven stylistic analysis because despite its novelty it has been thoroughly interpreted by critics due to its enormous popularity, which assigned it a status of a cultural phenomenon of early 2000's. With almost half-billion books sold, the novels are also interesting for their ability to capture a wide range of audiences, although they were originally written for pre-teen and/or adolescent readers. Moreover, the series has been analyzed from the point of view of various fields of linguistics, Diesing (2007) examined the act of utterance when casting spells and the notion of performativity of such, Brett (2009) explored translation of the dialectal speech in the series, and Nygren (2006) has completed a quantitative analysis on how the use of reporting verbs etc. works in the depiction of characters. It has been also touched upon by corpus linguistics in particular, for example by Čermáková and Fárová (2010) who examined the keywords in Harry Potter and their Czech and Finnish translation equivalents using a parallel corpus in an attempt to establish whether the translator is aware of the concept of keyness and how/whether this is reflected in translation.

4. Material and Methods

4.1 Corpora used for analysis

In the following analysis, the complete Harry Potter series of seven novels forms the study corpus (henceforth Potter) which was compared to a selection of texts extracted from the British National Corpus (henceforth teen-BNC).

4.1.1 The Harry Potter series

The Harry Potter series is a heptalogy of fantasy novels, the first one being released in 1997 with the six sequels coming out in the following decade. The initial books of the series were primarily considered to be works of children's literature but the closing part of the series was due to its darker tone and more complex development of plot extended over larger volume of text targeted on older audience of adolescents and young adults. The Potter corpus consisting of the seven novels was further divided into smaller subcorpora each consisting of an individual novel (henceforth HP1-HP7) (cf. Table 2 below).

Table 2: The corpora used in the research

corpora	words	
	types	tokens
HP1	5717	80635
HP2	6836	88545
HP3	7365	110914
HP4	10206	197228
HP5	12074	265884
HP6	10281	174816
HP7	11119	204344
Potter	21275	1122366
BNC	34977	1743590

4.1.2 The reference corpus

Because the 100-million British National Corpus composed of a variety of texts would be very likely to "absorb" or make the distinguishing stylistic features found in the Potter corpus statistically irrelevant simply due to it volume, some restrictions were needed when selecting the material for the reference corpus. In order to achieve greater

accuracy in the stylistic analysis, the textual material obtained from BNC was selected to match the Potter corpus in the textual type and genre, that is, only relevant textual samples of fiction works whose target audience is that of preteens and adolescents. Because most of the texts included in the BNC were published between 1975 and 1993, the body of the reference corpus is composed of texts slightly older than that of the Harry Potter series. The BNC texts were used because no newer collection of such literary works was available but it is assumed that the changes in both lexis and grammar are minor within this time period. Tenn-BNC includes all books for teens found in BNC but only 8 of these were published between 1960 and 1984, the rest was published between 1985 and 1993. This assumption should be especially valid for the written text as it does not undergo as rapid development as the spoken language and appears to have a higher level of grammatical regularity. Since only the texts of the American versions of the Harry Potter series were available and the reference corpus consists of books published in Britain, the possible influence of the differences between British and American English had to be considered when calculating the keywords. The divergences caused by the use of the American 'translation' for the primary sources should be however minor, especially because the extent of the translation decreased in the latter part of the series as it was targeted to older and supposedly more knowledgeable audience. Where relevant, the effects and drawbacks of these differences will be discussed below.

4.2 AntConc: The software and its functions

A freeware concordance software AntConc, version 3.2.2.1 (Anthony, 2011) was used to generate a keyword list. Since AntConc allows the user to set various parameters, several values were adjusted for the purpose of this research in order to, for example, increase the level of accuracy and avoid handling large sets of unecessarily detailed and extremely localized data with uneven dispersion throughout the corpus.

4.2.1 Using the Word List tool

AntConc contains seven major tools, out of which several were used in this research. First, word lists were compiled out of individual HP corpora using the reference corpus using the Word List tool which simply detects the items with the highest absolute frequency in an corpus. Such list may be ordered or inverted, arranged alphabetically based on the first or the last letter in a word but more importantly set to be case-insensitive mode, an option that was selected for this research to avoid potential bias where a split of the capitalized sentence-initial words and those written with lower-case letter.

4.2.2 Using the Keyword List tool

This step was followed by generating the keywords out of the respective data set, using the Keyword List tool, a process, in which AntConc "compares the words that appear in the target files with the words that appear in a 'reference corpus' to generate a list of

"keywords", that are unusually frequent (or infrequent) in the target files." (Anthony, 2011) The resultant words are the central items for this research will be subject to further analysis and interpretation. There was no need to set the minimum frequency cut-off point since we set out to explore only the one hundred highest-ranking keywords. For calculating the level of keyness and therefore relative significance of each word, log-likehood test was selected over Chi-square as it is the preferred research method in papers with comparable objective (Fischer-Starcke 2009, Culpeper 2009) and also an option recommended by the author of the software. The threshold value of keyness was set to 10.827; all the keywords in the set examined exceeded the critical value, and can therefore be considered highly statistically significant. The software also allows the user to make a choice whether only a single list of keywords is to be retrieved (positive keywords) or a list of negative keywords i.e. words with unusually low frequency in a given text, shall be generated as well. Because the rarity of certain elements may also efficiently serve as a style-marker, the setting that allows generating of both sets was selected. An example of such deviation in terms of negative frequency is e.g. numerously cited case of the character of Lok from the Inheritors whose language is most significantly marked by the fact that he does not use transitive verb-forms. (Leech&Short, 1981: 49)

4.2.3 Other tools

Apart from the functions mentioned above which should serve as principal tools in this reseach, AntConc also allows a close study of concordance lines of individual words which allows manual sorting of these and verification of words in their sentential context via the Concordance tool. The Concordance Plot tool has a similar function, offering an option of an alternative view of the concordance lines where all the occurrences of a word in a file are displayed in a liner sequence with wider vertical lines indicating more frequent use in a particular segment of the text. This allows the researches to observe the individual item's dispersion within the text, Furthermore, it can serve for retrieving ordered lists of clusters that appear in the proximity of the search term and can be also used "to identify where the search term hits cluster together." (Anthony) The Clusters tool enables the researcher to generate "an ordered list of clusters that appear around term in the target files" (Anthony). This function allows ordering of the search results similarly as e.g. that of the keyword listing, including their inversion. Moreover, limits can be set regarding the minimum and maximum length of the segment, it position in relation to the search term and also the cut-off frequency limit of the displayed items. (Anthony) Last but not least, the N-Grams tool generates word N-grams, repeated sequences of at least two words that appear in the target files.

4.2.4 Limitations due to the software and textual material

Due to the lack of lemmatization morpho-syntactic markup, i.e. tagging, the program is not aware of the words' lexical meaning in larger context and its word-class, and therefore does not recognize homographs as individual lexical items and merges the two

(or more) separate units together which ultimately leads to distortion of their actual levels of keyness. There is also a large number of English words whose morphological structure cannot serve as an indicator of the word-class. These often have an ability to appear as more than one word classes. Even though we can view keywords in context, the distinction between grammatical, lexical and other categories of words is sometimes blurry. This may be partly dealt with by manually sorting out each individual use in the concordance plot lines but establishing the actual levels of keyness for each item remains a problem. Since treat-all-data-as-lower-case option must be used in order to avoid distortion of data due to the words with capitalized initial letters at the beginning of each sentence yet another problem emerges: Rowling capitalizes a number of items when they are treated as a particular magical object (*the Sorcerer's Stone, the Mirror of Erised*) but these are also frequently used in their common meaning (*they clambered up the stone steps, he watched the cat in his mirror*) but both units ultimately end up merged together under a single search result with comparatively higher keyness level. A problem also occurs with some common last names that are homonymous to lexical words (*Wood, Black*) and outside of context only differentiate from their generic counterparts by the initial capital letter which unfortunately loses its significance with the program set on treat-all-data-as-lower-case option as the two inevitably merge. There is, however, a noteworthy number of words for both magical items and creatures (*wand, centaur, goblin*) that have previously appeared in various literary sources or folklore/oral traditions which are never capitalized unless at the beginning of a sentence. There are also several occurrences of a word being used idiomatically rather than in its primary meaning but these minor divergences will be ignored for the purpose of this paper.

4.3 Research method

Having compared the resultant word lists using the keyword list tool, the retrieved keywords will be split into two categories according to their textual function: The first group should encompass all the lexical words including the category of proper nouns and the second all of those whose function is more stylistic or grammatical. The results that fall into the first group will be further subdivided according to their lexical domains. The results that fall into the second group will be categorized according to their functional domains, thus separating the words whose function is merely grammatical from those that are more of stylistic features or function to establish interpersonal relations (cf. Halliday 1973,1978,1994, Culpeper 2009) in order to prepare the data for interpretation.

5. Keyword analysis of the Harry Potter series

This section presents the results of keyword analysis performed on the Harry Potter series. In the following lines, the classification of the obtained keywords (cf. Table 3) into two large groups encompassing grammatical and lexical words, proposed earlier in this paper, will be adopted. Its advantages, drawbacks and limitations will be discussed

in the concluding remarks along with a commentary on the results obtained using various tools of the AntConc program.

5.1 Grammatical and stylistic keywords

The grammatical keywords reflect on the way the text is organized but also serve to establish relationships between its individual actions and characters and may therefore provide very useful information regarding the plot. Even though it comprises mostly of grammatical words, this chapter also deals with discourse markers and other "indicators of the communicative purpose and micro- or macro- structure of the text" (Bondi & Scott, 2010: 7) Because they are virtually unpredictable, the grammatical keywords have a great potential to reveal subtle details that may significantly influence our reading of the text. The most common problem encountered when analyzing the grammatical keywords is the extremely time consuming processing of their concordance lines simply due to their number and also their frequent multiple uses, e.g. *who* may be used as either a relative or an interrogative pronoun, and *though* may function as a part of compound conjunctions, such as *even though* and *as though as*. Last but not least, their dispersion within the text is far more even than that of the lexical keywords which are more topic-bound and therefore somewhat more localized.

5.1.1 Pronouns: personal and possessive

Perhaps the most predictable grammatical items appearing on the keyword list are masculine pronouns, namely personal *he* and possessive *his*. These are quite predictable for several reasons. First of all, majority of the series' characters including the protagonist and the antagonist are male. Secondly, apart from occasional parts that take the form of a letter or a diary entry, and are therefore are written in the first person, the novel is a consistent third-person narrative so the first and second person pronouns are restricted to the special passages mentioned above and direct speech, whereas there are no constraints as to where the third person pronouns may appear. *Yeh*, a spelling variant of second person personal pronoun *you* only appears on the listing due to its non-standard orthography which serves to represent one of the character's, Hagrid's, accent and aside from functioning as a tool to portray the character's place of origin and lack of education is of very little structural importance. Third person pronouns have largely anaphoric function, referring to items that have previously appeared in the text. *He* and *him* therefore frequently have replacive, economic and cohesive function in the text, substituting previously mentioned common and proper nouns in objective and possessive case, respectively, in order to avoid time and space consuming repetition of the former.

Table 3: Top 100 keywords, ranked by keyness (log-likelihood)

rank	freq.	keyness	keyword
1.	18140	33122.139	harry
2.	6309	11828.908	ron
3.	5344	10019.605	hermione
4.	3344	6269.753	dumbledore
5.	2030	3806.100	hagrid
6.	2038	3641.113	professor
7.	1814	3401.116	snape
8.	1652	3056.487	wand
9.	1587	2975.508	weasley
10.	14448	2751.451	said
11.	1322	2478.652	malfoy
12.	1223	2293.035	voldemort
13.	1181	2126.270	potter
14.	1131	2120.541	sirius
15.	877	1644.310	hogwarts
16.	809	1516.815	lupin
17.	772	1447.443	ginny
18.	771	1445.568	neville
19.	910	1425.021	fred
20.	760	1378.922	toward
21.	733	1374.321	mcgonagall
22.	631	1183.078	umbridge
23.	607	1138.080	gryffindor
24.	14282	1112.223	his
25.	611	1053.106	ministry
26.	2192	996.720	around
27.	710	957.194	yeah
28.	1816	923.639	though
29.	496	905.889	moody
30.	470	881.215	dobby
31.	502	874.255	vernon
32.	478	872.288	fudge
33.	425	796.844	quidditch
34.	433	770.306	dudley
35.	509	762.240	wizard
36.	402	753.720	slughorn
37.	22224	710.833	he
38.	476	694.577	robes
39.	776	681.447	george
40.	460	680.636	yeh
41.	410	677.758	percy
42.	373	676.411	ter
43.	355	665.599	eaters
44.	353	661.849	slytherin
45.	1785	653.949	looking
46.	348	652.474	muggle
47.	343	643.100	luna
48.	381	631.221	students
49.	538	608.859	cloak
50.	315	590.602	cedric
51.	312	584.977	kreacher
52.	354	582.049	wizards
53.	14337	574.269	s
54.	287	538.104	petunia
55.	298	520.164	riddle
56.	276	517.480	filch
57.	333	515.591	crouch
58.	272	509.980	dementors
59.	383	504.460	magical
60.	253	474.356	krum
61.	250	468.732	lockhart
62.	243	455.607	fleur
63.	3306	455.586	who
64.	242	453.732	tonks
65.	235	440.608	goyle
66.	260	436.080	parchment
67.	232	434.983	crabbe
68.	231	433.108	bagman
69.	224	419.983	dursleys
70.	279	414.631	scar
71.	10113	411.755	had
72.	620	410.021	magic
73.	286	408.330	potion
74.	336	397.986	madam
75.	212	397.484	cho
76.	204	382.485	hedwig
77.	199	373.110	bellatrix
78.	561	365.675	upon
79.	195	365.611	wormtail
80.	194	363.736	wands
81.	190	356.236	albus
82.	430	346.386	slightly
83.	180	337.487	azkaban
84.	235	337.461	draco
85.	304	337.277	bill
86.	205	335.774	invisibility
87.	210	333.102	dean
88.	175	328.112	muggles
89.	210	327.601	grounds
90.	248	326.486	trelawney
91.	180	326.089	seamus
92.	328	321.894	er
93.	624	321.286	hall
94.	430	315.347	onto
95.	173	313.044	snitch
96.	166	311.238	defense
97.	278	311.123	okay
98.	163	305.613	peeves
99.	8670	304.997	at
100.	446	304.108	office

Table 4: Grammatical keywords within the top 100 keywords, ranked by keyness (Log-likelihood)

rank	freq.	keyness	keyword
20.	760	1378.922	toward
24.	14282	1112.223	his
26.	2192	996.720	around
27.	710	957.194	yeah
28.	1816	923.639	though
37.	22224	710.833	he
40.	460	680.636	yeh
42.	373	676.411	ter
53.	14337	574.269	s
63.	3306	455.586	who
71.	10113	411.755	had
78.	561	365.675	upon
92.	328	321.894	er
94.	430	315.347	onto
97.	278	311.123	okay
99.	8670	304.997	at

Table 5: Grammatical keywords within the top 100 keywords divided by functional domains

functional domain	keywords	number of items
prepositions - spatial arrangement, direction	toward, around, upon, onto, at, ter*	6
particles - agreement	okay, yeah	2
interjection - hesitation	er	1
pronouns – personal, posessive	he, his, yeh	3
pronouns – relative, interrogative	who	1
subordinator	though	1
adnominal case marker, marker in contractions	s*	1
past perfect marker, causative use	had*	1
total number		16

* used also as an infinitive marker

5.1.1.1 Personal pronoun *he*

Anaphoric personal masculine pronoun *he* is the fourth most frequent word in the Series (22224 concordance lines). In order to retrieve recurrent patterns which may enable us to better understand how *he* is used in the novel and why it occurs on the Top-100 keyword list (cf. Table 3), the Clusters tool will be used. One of the obvious benefits of using the Clusters tool is that it allows the researcher to "shrink" thousands of concordance lines into few dozens of the most salient lexical structures. The search settings were set to the clusters sized between five and six words with minimum of ten occurrences (cf. Table 6).

Table 6: Clusters around *he*

freq.	cluster
48	he-who-must-not-be
48	he-who-must-not-be-named
33	he did not want to

18	as fast as he could
18	he didn't want to
17	as though he had been
17	he felt as though he
16	he did not know what
16	looked as though he was
15	he got to his feet
15	i don't think he
14	he was supposed to be
13	as hard as he could
13	harry felt as though he
13	looked as though he had
12	he would be able to
11	as though he had just
11	before he could stop himself
10	all he knew was that
10	he's going to be
10	what he was going to

Perhaps the most surprising is the fact that in the most recurrent clusters *he* does not refer to the protagonist but the antagonist. *He-who-must-not-be-named* is a replacement euphemism for the tabooed name *Voldemort* that is capitalized in the series; we may therefore view it as a proper name or a nickname. Although not as numerous in their occurrence, most of the other clusters center around the protagonist: There is a focus on his mental processes, be it information processing (*he did not know that, all he knew was that*) feelings (*he felt as though he, Harry felt as though he*), wishes (*he did not want to, he didn't want to*) and evaluation of future prospects and plans, often hypothetical (*he would be able to, he's going to be, what he was going to*). This density of clusters denoting mental processes stands somewhat in opposition to the strong focus on the physical setting and lively action of the Series; they suggest that certain part of the novels does not take place in the fictional world but in the mind of the protagonist which the limited omniscient third-person narrator deliberately exposes to the reader, allowing him or her to keep track of Harry's train of thought. We may assume that the series retains its focus on the physical action but due to the author's use of more diverse vocabulary, the words do not become key. Some of the clusters also comment on the concord (or clash) between the reality and appearance (*as though as he had been, looked as though he was, he was supposed to be, as though he had just*) or express degree by comparison (*as fast as he could, as hard as he could*) (cf. Example 1).

Example 1: Selected concordance lines with clusters around *he*

*He was pale and **looked as though he was** about to cry. /HP1/*

*Uncle Vernon and Aunt Petunia froze when they stood staring at Dudley **as though he had just** expressed a desire to become a ballerina. /HP7/*

*He pondered for a moment then set off again, eyes closed, concentrating **as hard as he could**. /HP6/*

This is due to the narrators' premeditated subjectivity, even though the narrator is omniscient (though very selective in what scenes will be revealed and in what sequence) the emotional and attitudinal tone of the novel is strictly based on perception of the protagonist, we view the plot through Harry's eyes. Not fully acquainted with the world

of wizards, he inevitably compares unknown magical objects, creatures and actions to like things from the Muggle world that he is very well familiar with.

5.1.1.2 Possessive pronoun *his*

Because there are thousands of concordance lines for *his* (14282 occurrences), the most suitable methods of investigation are again using the Collocates Tool and the Clusters Tool, rather than the manual analysis of individual lines. We set the collocates' window to include two words to the right of the search term *his* in order to explore the nouns which function as heads of the noun phrases determined by *his*, whether premodified or not. The nouns appearing in the first hundred R1-2 collocates of *his* are listed in Table 7 bellow:

Table 7: Nouns appearing in 100 R1-2 collocates of *his*

rank	freq.	collocate
1.	751	wand
2.	741	eyes
3.	624	head
4.	583	face
5.	445	hand
6.	354	feet
7.	319	voice
8.	239	hands
9.	222	mouth
10.	180	robes
11.	168	mind
12.	152	hair
13.	139	shoulder
14.	138	father
15.	133	mother
16.	132	heart
17.	129	nose
18.	126	scar
19.	125	fingers
20.	121	arm
21.	117	arms
22.	113	pocket
23.	109	chest
24.	105	back
25.	106	way
26.	106	eye
27.	103	cloak
28.	101	life
29.	98	glasses
30.	98	ears
31.	96	body
32.	95	bed
33.	94	throat
34.	92	stomach
35.	90	trunk
36.	89	knees
37.	86	forehead
38.	85	neck
39.	79	chair
40.	73	legs
41.	72	parents
42.	72	name

The prominent semantic domains appearing on the list are body parts (*eyes, head, face, hand, feet, hands, mouth, hair, shoulder, heart, nose, fingers, arm, arms, chest, back, ears, body, throat, stomach, knees, forehead, neck, legs*), family members (*father, mother, parents*), magical objects (*wand, robes, cloak*) and furniture (*bed, chair*). First of all, vast majority of these nouns are concrete, physical objects, the only exceptions being *mind, life, way* and *name*. These "exceptions" can be better explained looking back at the clusters including *he.* Similarly, these clusters including *mind* often denote mental processes, thinking something over etc. (*in his mind's eye, out of his mind*) (cf. Appendix 1). Those including *life* often function as a part of comparisons or as a way of presenting some unique or surprising object or experience (*had ever seen in his life*) or an intensifier (cf. Appendix 2) (cf. Example 2).

Example 2: Selected concordance lines with clusters around *his*

Even as he said it, Harry remembered that his father had been pure-blood, but he pushed the thought ***out of his mind****; he would worry about that later. /HP 6/*

Harry rolled over in bed, a series of dazzling new pictures forming ***in his mind' s eye****. /HP4/*

She was carrying the largest block of chocolate ***he had ever seen in his life****. /HP3/*

The tremendous number of body parts appearing on the list is partly due to the nature of the English language, which requires a possessive determiner (unlike e.g. Czech, German or Spanish) when speaking about a person's body parts but more importantly, this discovery enables us to actually see that the numerous mentions of individual body parts which would otherwise remain obscured since the semantic domain is composed of various words which do not appear as key separately. The frequent occurrence of the body parts suggests certain focus on the physical action or simply physical being of the character. Perhaps more interesting is *his* collocating with *way*, which is often used in the metaphorical sense in idiomatic expressions such as *make/cut/bang/edge/claw his way to/into/towards*. These expressions show high degree of flexibility and are typical of informal, colloquial language.

Another method that may be helpful in revealing how *his* is used in the Series is using the Clusters Tool. Clusters composed of 5-6 items surrounding *his* were generated using the Clusters Tool (cf. Table 8).

Table 8: Clusters including *his* sized 5-6, minimum occurrence 10

freq.	**cluster**
19	harry got to his feet
19	on the back of his
18	of the corner of his
18	out of the corner of his
17	got to his feet and
16	over the top of his
15	he got to his feet
14	out of his pocket and
13	pointed his wand at the
13	the back of his hand

12	in the pit of his
12	pointing his wand at the
12	the pit of his stomach
12	the tip of his wand
12	the top of his head
11	first time in his life
11	his copy of advanced potion
11	his copy of advanced potion-making
11	in the pit of his stomach
11	the back of his head
11	the end of his wand
11	the first time in his
10	at the top of his
10	for the first time in his
10	of the corner of his eye
10	the corner of his eye
10	the first time in his life
10	the pain in his scar
10	the scar on his forehead
10	turned on his heel and

Glancing at the table, we may observe that *his* is frequently a part of sequences that express some kind of movement, as in *he got to his feet, pointed his wand at the, turned on his heel and*, etc. Other clusters often specify a part of a physical object (*the tip of his wand, the end of his wand* etc.) or a part of a body part or limb *(the back of his hand, in the pit of his stomach, the back of his head , of the corner of his eye* etc.). The author therefore makes subtle distinction between individual components of objects, indicating the precise part that is involved in the action. Others specify direction, origin (e.g. *out of his pocket*), location (e.g. *the scar on his forehead*) or denote magical objects (e.g. *his copy of advanced potion-making*). The clusters including *life* (*e.g. the first time in his life*) have, as was commented on in previous lines, more of an intensifying function, intriguing the reader by presenting some very unique, unprecedented phenomenon. Last but not least, some of the clusters mentioned above help to expose some recurrent verbs which would otherwise remain concealed, such as *point, get* and *turn*, suggesting that one of the reasons for such a small representation of the word class on the keyword list may be Rowling's choice of rather simplistic vocabulary when it comes to action verbs.

5.1.2 Interrogative and relative pronoun *who*

Due to the large number of concordance hits (3306), it is impossible to manually (or quantitatively) detect the prevalent use of the pronoun *who*, which may appear as either an interrogative or a relative pronoun. Based on a manual overview of the first 100 resultant concordance lines and general assumptions about the functional structure of English language it is, however, more plausible that the relative use prevails and therefore suggests frequent use of subordinate relative clauses which may be vaguely classified as providers of additional information to the content of the main clause. The overviewed concordance lines exhibited mainly relative use (68/100), followed by the nickname of the antagonist (You-*Know-Who, He-Who-Must-Not-Be-Named* – 20/100) and interrogative use (10/100). The remaining two uses are part of a nickname used for the protagonist (*The Boy Who Lived*).

Consistent use of subordination is nevertheless one of the principal features distinguishing the written text from the spoken word and therefore not anything uncommon in general fiction. The reason for the prominence of *who* is thus likely to lie elsewhere: Interestingly enough, Rowling capitalizes *who* when it functions as a part of proper names or nicknames such as *The Boy Who Lived* or *He-Who-Must-Not-Be-Named* (also *You-Know-Who*). When generating a list of clusters around *who* sized 4-5, minimum of occurrences 10, the vast majority is formed precisely by these nicknames (cf. Appendix 3), followed by proper names that suggests the author's frequent use of non-restrictive relative clauses, functioning as postmodification (cf. Appendix 4). The proper names including *who* are not infrequent and therefore inevitably distort the actual number of occurrences of the standard use of relative and interrogative *who* to such an extent that without the author using *who* as a part of proper names, it would most likely not appear on the keyword list.

5.1.3 Prepositions: spatial arrangement and direction

A prominent semantic domain appearing on the list of the lexical words that consists of items denoting spatial arrangement has its grammatical counterpart: there are multiple prepositions of space (cf. Table 5). This is only logical since the prepositions, being synsemantics (Dušková et al., 1988: 273) are only capable of forming a clause element together with the noun their accompany, expressing the relationship between a noun and another noun, a noun and a verb, or an adjective an a noun (cf. Dušková, 1988). Although the majority of the most common prepositions is monosyllabic and carries no stress in speech, the keywords retrieved from the Potter corpus are mostly polysyllabic and therefore those that would normally receive stress in speech. Interestingly enough, many of them are also structurally compound unites of the adverb + preposition type. It must be, however, noted that two items with the highest level of keyness out of these only appear on the keyword list due to differences between American and British English: American English speakers show preference for using *toward* where British English would use *towards*[2] and similarly American English allows the use of *around* where British English speakers opt for *round*[3].

5.1.3.1 Further implications of the choice of prepositions

The occurrence of these prepositions along with their lexical counterparts i.e. nouns denoting spatial locations (*i.e. grounds, hall, office*…) among the highest-ranking keywords suggest that the Series' plot and development of individual actions has very definite setting in space. The choice of compound and lexically more specific prepositions over primary prepositions (Dušková et al., 1988: 275) that have „more general meaning and some of which have been weakened to mere indicators of case relationships" (ibid) again points towards the author's sensitivity to arrangement of objects, people an action in space. Another point that is to be made is that the vast majority of the prepositional keywords may be classified as dynamic (cf. Dušková et al.,

[2] http://www.oxfordadvancedlearnersdictionary.com/dictionary/towards
[3] http://www.oxfordadvancedlearnersdictionary.com/dictionary/around

1988) i.e. *where from, which way, where to,* as opposed to the static prepositions – *where at*. Last but not least it is important to note that the *ter* (a variant of spelling of *to* used to represent a character's accent and/or non-standard pronunciation) is only in included because of the way it is written, just like several other misspelled words used to represent Hagrid's idiolect. Its importance therefore lies in the way it characterizes one of the protagonist by non-standard spelling rather than its identity as a part of speech.

5.1.3.2 The use of prepositions in the construction of the fictional landscape

The author offers very precise and detailed descriptions of spatial arrangement of the characters whereabouts, therefore allowing the reader to visualize the fictional landscape more easily. Grouped with other words, they function cohesively. The authors of CGEL elaborate on their function as a part of place relators, saying that normally "shifts or distinctions within space tend to be expressed by location-specific lexical items or even proper noun (place names for example)" (Quirk et. al, 1989: 1448), while admitting that "certain spatial relations are firmly linked to grammatical expressions which are heavily exploited in textual structure." (ibid) They further comment on the structural makeup of the place relators suggesting that "place relators often comprise two compounds. Most frequently these are dimension or direction indicator plus a location indicator. The latter is usually an open class noun (or proper noun) but its locational use is often institutionalized, making the whole expression quasi-grammatical." (ibid, 1449)

As we may observe in the concordance lines listed bellow (cf. Example 3), the prepositions may be either independent or form a part of prepositional verbs such as in *bump into*. This makes, however, only a very little difference as the semantic content of the prepositional verbs remains rather similar to the sense of the original components. The author not only provides static description of the physical surroundings but also very carefully works with directions (e.g. movement in space but also where the light comes from, where shadows are being cast), thus creating a truly three-dimensional space.

Example 3: Selected concordance lines with key prepositions

*At this moment the boat bumped gently **into** the harbor wall. Hagrid folded up his newspaper, and they clambered **up** the stone steps **onto** the street. /HP1/*

*The narrow path had opened suddenly **onto** the edge of a great black lake. Perched **atop** a high mountain **on** the other side, its windows sparkling **in** the starry sky, was a vast castle with many turrets and towers. /HP1/*

*Evil-looking masks stared down **from** the walls, an assortment of human bones lay **upon** the counter, and rusty, spiked instruments hung **from** the ceiling. /HP2/*

*He lifted the basin, carried it **over to** his desk, placed it **upon** the polished top, and sat **down in** the chair **behind** it. He motioned for Harry to sit **down** opposite **him**. /HP4/*

*A thousand flecks of golden light sparkled **upon** the dark surface of the water a few feet **below** where he crouched; the black wall of rock **beside** him was illuminated too. /HP7/*

The prepositional phrases comprising of the complex prepositions typically function as adverbials. The noun phrases complementing the prepositions frequently contain premodification (e.g. *harbour wall, stone steps, high mountain, starry sky, polished top, dark surface*) unlike the noun phrases with the determiner *his*, thus making the portrayal of fictional scenery even more vivid and concrete.

In order to investigate the organizational structure of the text more thoroughly, we will look now at the individual preposition's collocations. Since the two items with the highest keyness level only appear on the list due to difference between American and British English and the third highest ranking item is only present due to its unconventional spelling, they will not be discussed.

5.1.3.3 Collocates of *upon*

When researching the most frequent collocates within two words to the right (cf. Appendix 5), that is, the words that are likely to be a part of the prepositional phrase, the preposition's most frequent collocates apart from strictly structural words are lexical units denoting people or some sort of a human presence such as *him, Harry, his, her, Dumbledore* with the words denoting physical items or things appearing only somewhat later and in lesser numbers on the list, e.g. *floor, table, desk, door*. Because *his* may serve as a determiner to an inanimate object as well, the immediately following collocate of *upon his* was generated, the highest ranking items being *chest, face, head, ears* and *back,* therefore confirming the previous hypothesis about human presence. This suggests that not only physical objects but also human beings are put very directly and precisely in certain position (be it spatially or socially) or relation towards each other. When looking at the 1-2L collocates of *upon* (cf. Appendix 6), the most frequent word is *down*, as in *down upon,* which is most often preceded by the verbs *bear, look* etc. (cf. Appendix 7); several other expressions denoting subjective perception of the story through the eyes of the protagonist appear.

5.1.3.4 Collocates of *onto*

Whereas *upon* was more likely to put humans into certain spatial or social context or position, the collocates of *onto* are rather different. Within 1R-2R, *onto* collocates (putting aside the strictly structural words) with a noun phrase whose head is frequently formed by larger physical objects denoting either furniture or some type of a surface, such as *floor, field, bed, grounds, platform, ground, table, grass* (cf. Appendix 8). There is, however, a slight exception to the rule: There are four high ranking collocates denoting human presence *his, Harry* and *back*, suggesting that *onto* may be used also to refer to actions performed with a part of one's body such as in *the goblin clambered onto his back.* Perhaps of more interest are the immediate left collocates of *onto,* where a verb would normally be expected. The most frequent lexical subgroup directly preceding *onto* is nevertheless that of adverbs denoting direction such as *back, out,*

down, up, backward, over forward (cf. Appendix 9). Presumably, these are employed to provide even more specific (cf. also *down upon*) and exact location of an object or an action as in *he had to flop back down onto the grass*.

5.1.3.5. Collocates of *at*

Another high frequency preposition, *at* has quite predictable immediately preceding collocates such as the verbs of vision, pointing etc. (*looking, looked, staring*, *stared, pointing, glanced*) but similarly to *onto*, it also collocates with adverbs of directing that specify the target location eg. *up, down, around, back, out, ever* such as in *Mrs.Weasley beamed down at him* (cf. Appendix 10). It's collocates to the right do not show any uncommon tendencies, they include parts of noun phrases (*the, him, Harry, his, Hogwarts, her*) that frequently form syntactic object and words that are parts of fixed expressions (*at once, all, last, least*).

5.1.4 *Had* – temporal relator

With this rich supporting structure of complex place relating structures we must inevitably wonder at the apparent lack of the time relators in the list of top-ranking keywords. The only plausible time relator that appears on the keyword list is *had*, some of its most frequent collocates immediately to the right being *been, come, gone, seen, happened, done* and *taken*, which are clear indicators of the past perfect, suggesting that the plot has complex temporal perspective where the particular sequence in which the actions occurred is carefully distinguished. *Had* therefore represents an important temporal tie, suggesting that the author turns to temporal system as means of establishing time relations more often than other authors of children's literature. The other time relators do no exhibit any special tendencies; we may therefore presume that their use does not differ from other works of children's literature. *Had* may also serve as a past periphrastic form of modal verb. Frequently collocating at 1R with *had*, an infinitive marker to *to* indicates the modal use. That is, however, largely outnumbered by the past perfect verb forms.

5.1.5 Particles denoting agreement

Okay and *yeah* are the most basic means of verbal expression of agreement. They are almost exclusively used in direct speech and their secondary function is to inform the speaker that the listener is keeping the track with what is being said. Their frequent use suggests colloquial dialogical direct speech where information is being exchanged between one or more characters. This assumption is further confirmed by the lexical keyword *said* which has a second highest number of occurrences after the first name of the protagonist.

5.1.6 Interjection denoting hesitation

Er is used to represent a hesitation or a hesitant pause in speech where the speaker makes an attempt to gain time to prepare his or her further lines. When used in writing

in direct speech it helps the text to resemble human speech more realistically. The author's frequent employment of this interjection points to her sensitivity to the actual prosody of human speech and emotional undertone of what is being said, thus making the dialogue more credible.

5.1.7 Conjunction of concession

Though appears in the text as an individual conjunction, a simple subordinator of concession (cf. Quirk et al;, 1989), but also as a part of complex subordinator in adjuncts of comparison *as though as* and *even though,* complex subordinator of concession. When used as a conjunction on its own, it is considered to be less formal and more colloquial than *although*; the high occurrence of this form may be therefore again connected with frequent occurrence of dialogical direct speech.

5.2 Lexical keywords

Among the top 100 keywords generated by comparing the Potter corpora to teen-BNC there are 28 lexical keywords that are not proper names (cf. Table 9). For the purposes of further research, these may be efficiently divided according to the semantic domains they represent (cf. table 10).

Table 9: Lexical keywords (excluding names of characters) within the top 100 keywords, ranked by keyness (log-likelihood)

rank	freq.	keyness	keyword
6.	2038	3641.113	professor
8.	1652	3056.487	wand
10.	14448	2751.451	said
25.	611	1053.106	ministry
33.	425	796.844	quidditch
35.	509	762.240	wizard
38.	476	694.577	robes
43.	355	665.599	eaters
45.	1785	653.949	looking
46.	348	652.474	muggle
48.	381	631.221	students
49.	538	608.859	cloak
52.	354	582.049	wizards
58.	272	509.980	dementors
59.	383	504.460	magical
66.	260	436.080	parchment
70.	279	414.631	scar
71.	10113	411.755	had
72.	620	410.021	magic
73.	286	408.330	potion
74.	336	397.986	madam
80.	194	363.736	wands
82.	430	346.386	slightly
86.	205	335.774	invisibility
89.	210	327.601	grounds
93.	624	321.286	hall
95.	173	313.044	snitch
96.	166	311.238	defense
100.	446	304.108	office

Table 10: Lexical keywords divided by the semantic domains they represent

semantic domain	Keywords	Number of items
space, place	ministry, grounds, hall, office	4
people, persons - general	professor, wizard, students, wizards, madam	5
action/sensory/state verbs	said, looking, had*	3
adverbs - degree	slightly	1
magical beings	Death Eaters, Dementors, Muggle	3
magical items, concepts	wand, Quidditch, robes, invisibility, cloak (as in the Invisibility Cloak), magical, parchment, scar, magic, potion, wands, Snitch, defense (as in the Defense Against Dark Arts)	12
Total number		28

*Had has been included both among grammatical and lexical keywords since it can be either used as an auxiliary or as a lexical verb.

5.2.1 General animate nouns – persons, people

To begin with the most predictable semantic domain, the top 100 keyword list includes several nouns denoting persons with somewhat generalized reference. They are perhaps not general nouns in the sense of being on the borderline between grammatical and lexical cohesion (cf. Halliday and Hasan, 1976) but they denote very general entities: In the magical world of Harry Potter, the noun *wizards* becomes frequently synonymous to people, being an umbrella term for the fictional wizarding society (cf. Example 4). These nouns are both setting-specific (*students, professor, madam*) that reflect on the plot being set at a school where some formal student-teacher relations exist and genre-specific (*wizard, wizards*), reflecting on the Series being of the fantasy genre.

Example 4: Selected concordance lines with keywords *wizard* and *wizards* used with generalized reference

*As you know, underage **wizards** are not permitted to perform spells outside school, and further spellwork on your part may lead to expulsion from said school. /HP2/*

*These plants are most efficacious in the inflaming of the brain, and are therefore much used in Confusing and Befuddlement Draughts, where the **wizard** is desirous of producing hot-headedness and recklessness. /HP5/*

5.2.2 Inanimate nouns – magical objects, concepts

Similarly, words that denote magical objects and concepts are also a highly predictable set of keywords, knowing the genre and the general topic of the Series. If we are to take a closer look at these, we realize that almost all of these objects (apart from *robes* and *Defense Against Dark Arts* which is a name of a subject taught in Hogwarts) have an intrinsic ability to or are used to perform magic. That there are indeed, objects, creatures and persons capable of performing magic is further proved by the only high ranking adjective that is listed along with these for practical purposes – *magical*. Aside from

being a part of various proper names (e.g. *Care of Magical Creatures, St. Mungo's Hospital for Magical Maladies and Injuries*), this adjective often functions to simply distinguish between a generic real-world item and those tied to the fictional world that receive such premodification (cf. Example 5)

Example 5: Differentiating real world and wizarding world objects – the premodifier *magical*

Apparently Moody's ***magical eye*** *could see through solid wood, as well as out of the back of his head. /H4/*

A few people laughed; Harry caught Seamus's ***eye****, and Seamus winked./HP1/*

5.2.3 Nouns denoting place, location

Somewhat less predictable but very prominent group of lexical keywords is composed of the nouns denoting space, place or a location: *Ministry, grounds, hall* and *office*. Both ministry and office allow metonymical use, where the words refer to the individual institution rather than a physical building or its part. Even though these uses appear *I'd be kicked out of office for suggesting it* or *if the Ministry thinks it appropriate,* they are infrequent. Similarly, the word *office* often appears as a part of various multiword proper names such as in *Improper Use of Magic Office* or *The Misuse of Muggle Artifacts Office* but these occurrences are not too frequent either.

The author's sensitivity to subtle distinctions regarding place, direction and spatial organization of the fictional landscape has been extensively commented on in the preceding chapter on grammatical words. Frequently grouped with prepositions, these words often function as place relators along with other means of grammatical cohesion aid the reader's orientation within the textual structure.

5.2.4 Nouns denoting magical beings

Death Eaters, Dementors and *Muggles* are all nouns referring to the members of highly specialized groups of beings or creatures, the first standing for the followers of the Lord Voldemort, the second for the creatures guarding the prison of Azkaban and the last for the persons that do not possess the ability to perform magic. They were not listed along with proper names because they refer to a social class or type of beings and they all are frequently preceded by *a* when used in singular. Moreover, even though they are capitalized in the English version, other language versions, such as Czech, do not capitalize them, suggesting that they are understood to denote general entities rather than being proper names.

5.2.5 Adverb *slightly*

The only key adverb is a word used to express degree *slightly*. Looking at its most frequent collocate immediately to the right, *slightly* is most frequently combined with *as* (cf. Example 6).

Example 6: Selected concordance lines showing the use of *slightly* and its collocate *as*

Wood shot toward the ground, landing rather harder than he meant to in his anger, staggering ***slightly as*** *he dismounted.* /HP2/

Everyone drew back ***slightly a****s Hagrid reached them and tethered the creatures to the fence.* /HP3/

I***n the silence Harry imagined he could hear the snake hissing*** ***slightly as*** *it coiled and uncoiled--or was it Voldemort's sibilant sigh lingering on the air?* /HP7/

The construction of the cluster is typically participle + *slightly* + *as*-clause (finite). The use of the participial form has economical reasons, the non-finite verb form is used for condensing the background action whereas the following finite form expresses the main action. The participial clause is used to express (and background) accompanying circumstances. The second most frequent collocate of *slightly – and* - suggests multiple actions (cf. Appendix 11). The third most frequent collocate immediately to the right of *slightly* is *to*, a construction frequently used to modify actions and/or establish precise direction in space (cf. Example 7).

Example 7: Selected concordance lines with *slightly* used to modify direction

Harry had heard Fred and George Weasley complain about the school brooms, saying that some of them started to vibrate if you flew too high, or always flew ***slightly to the lef****t.* /HP1/

Harry, moving ***slightly to his right****, saw that Trelawney's terrifying vision was nothing other than Professor Umbridge.* /HP5/

"For the same reason that Avery, Yaxley, the Carrows, Greyback, Lucius"— he inclined his head ***slightly to Narcissa*** *— "and many others did not attempt to find him."* /HP6/

The other uses of *slightly* are diverse but it can be assumed that its use suggests the author's ability to make distinction between various degrees of a quality or an action even there where only a subtle differences appear.

5.2.6 Verb *said*

In a past-tense third person narrative, *said* is a word that we can safely predict to appear very high on the word list. When compared with the reference corpora composed of other narratives (not written exclusively in third person past tense though), it is rather striking for the word to have this high value of keyness. The reason behind is extraordinary frequency is the rate of the occurrence of dialogical speech. A direct dialogical speech in such narrative always sooner or later requires reporting verbs to comment on the fact that the words are being said, shouted or perhaps whispered by a certain speaker. The assumption about the frequency and therefore also the importance of direct dialogical speech can be further supported by high ranking keywords that are speech-bound – hesitation marker *er* and particles denoting agreement *okay* and *yeah* that appear exclusively in spoken lines. When researching the collocates of *said* immediately to the left, more spoken-language specific particles appear - *no, yes, right, yeah, well, okay*, confirming the hypothesis about lively dialogue being a vital part of the narrative (cf. Example 8).

Example 8: Selected concordance lines with *said* collocating with spoken-language specific particles

***"Yes," said** Quirrell idly, walking around the mirror to look at the back. /HP1/*

*"Yes, yes, that's **right," said** Professor Flitwick, beaming at Harry. /HP1/*

*"**Well, well, well!" said** Umbridge, looking triumphant. /HP5/*

5.2.7 Verb *looking*

Looking, the only key verb to appear in the *–ing* form is used in several ways. Because the corpus is not tagged and there were 1785 concordance lines generated, a method of sampling was selected for making further conclusion about the nature of the use of the word *looking*. The first twenty concordance lines from each part of the Series were analyzed. As it is typical of the participial clauses, the nonfinite verb form is very frequently used to present the background action, whereas the finite form expresses the action that is in the focus and carries more communicative dynamism.

Example 9: Selected concordance lines with *looking* used to present background action/state

*"Exactly." said Dumbledore**, looking** very seriously over the top of his half-moon glasses.* /HP1/

*"Hope you have — er — a good holiday," said Hermione**, looking** uncertainly after Uncle Vernon, shocked that anyone could be so unpleasant.* /HP2/

*Harry found this such an inadequate response to everything that had happened tonight that he turned the piece of parchment over**, looking** for the rest of the letter, but there was nothing else.* /HP5/

As we can see from the examples above, the use of the verb *looking* is not restricted to its primary meaning – a lexical verb of perception but it is also used in its secondary sense such as a copular verb and also as a part of various phrasal and prepositional verbs (*looking for, looking after, looking forward to, look around, look out* etc.) which are nevertheless derived from the primary form. Be it its primary sense or the additional one ascribing the subject some characteristics, the participial form very frequently collocates within the distance of two words to the left with the verb of speaking – *said*. The author therefore tends to frequently provide additional information to the way the characters look at (or avoid to do so, their gaze being directed elsewhere) the receiver of what is being said, what they look like while they say it or what facial expression they have and also when they are potentially doing while they speak. Because the five most frequent collocates of *looking* immediately to the right are *at, for, around, up, down* which when combined compose about a third of all occurrences (cf. Appendix 12), we may safely assume that *looking* is used most frequently in its primary lexical sense or as a phrasal verb *looking for*, followed by the copular use ascribing the subject certain characteristics:

Example 10: Selected concordance lines with *looking* used as a copular verb

He was ripping the paper off a gold wristwatch when Aunt Petunia came back from the telephone ***looking*** *both angry and worried. /HP1/*

Next moment Dudley came flying into the hall, ***looking*** *terrified. /HP4/*

Furthermore, Fudge was ***looking*** *distinctly careworn. /HP6/*

Looking has certain characterizing function also when employed as a part of compound premodifiers such as *funny-looking people, ruffled-looking owl, strange-looking coins, battered looking coins* or *decrepit-looking bird*. Attaching *looking* to adjectives is a lexical tool helps the reader to perceive the fictional world through the eyes of the protagonist: Harry has just entered the magical world for the first time in his life and is therefore not able to identify objects and people by their names or function but is able to judge them by their physical appearance. When analyzing the selected samples, only few isolated gerundial uses of looking were detected, such as in

Example 11: Selected concordance lines with gerundial uses of *looking*

"Ready, Duddy?" asked Petunia, fussily checking the clasp of her handbag so as to avoid ***looking*** *at Harry altogether*. /HP 7/

Ron, however, walked right past Hermione without ***looking*** *at her.* /HP4/

These were, however, not numerous. As for the fact that the fifth sequel of the series contains about a third of all the occurrences of *looking*, no other plausible explanation than the novel's voluminosity presented itself.

5.2.8 What did not appear on the keyword list

If we leave out the only adjective (*magical*) and adverb (*slightly*) and the three lexical verbs, no members of these world classes appear on the top-100 keyword list. It is very difficult to determine the cause of this, a possible explanation could lie simply in the fact the high ranking positions are occupied by the proper names and genre-specific nouns. Rowling invents a plethora of nouns or resurrects those found in mythologies, folk tales etc. but does not do so to such an extent with other word classes. She uses common English verbs for example, making them magical only by attaching a new meaning to them such as in *stun* (to knock one down using a spell) or *disarm* (to rid one of his or her wand using a spell) so they are not as likely to appear as key. As for the adjectives and adverbs, immense richness of her vocabulary a frequent use of synonymous terms in order to avoid repetition is likely to decrease the individual words potential to become key. If we look further down the list and extend the list by another hundred, we encounter somewhat less proper names and more variety of word classes including a handful of verbs in –ing and –ed forms (cf. Appendix 12). The verbs appearing in the second hundred exhibit semantic similarity to those appearing in top 100, thus confirming the previously stated hypotheses, they frequently denote communication (*asked, told, tell, say*), perception (*look, saw, heard*), cognition and feelings (*knew, felt*) and movement (*going, go, turned*). The nouns denote body parts (*face, voice, head, hand*) and places (*room, door*), the adjectives are rather simple, often

denoting two polar opposites on a scale (*good, long, dark, little, black, great etc.*) The most surprising semantic domain are the words expressing uncertainty and indefiniteness – *something, seemed, some, might*. These are very common words in any text type, it is therefore very interesting that they are so salient in the Series.

5.3 Proper names

The bulk of high ranking proper names (cf. Table 11) appearing on the keyword list was largely predictable as they correspond with the names of the protagonists, but what may be of more interest is the form they take and the way they can be accordingly classified (cf. Table 12).

Table 11: Proper names within the top 100 keywords ranked by keyness, (log-likelihood)

rank	freq.	keyness	keyword
1.	18140	33122.139	harry
2.	6309	11828.908	ron
3.	5344	10019.605	hermione
4.	3344	6269.753	dumbledore
5.	2030	3806.100	hagrid
7.	1814	3401.116	snape
9.	1587	2975.508	weasley
11.	1322	2478.652	malfoy
12.	1223	2293.035	voldemort
13.	1181	2126.270	potter
14.	1131	2120.541	Sirius
15.	877	1644.310	hogwarts
16.	809	1516.815	lupin
17.	772	1447.443	ginny
18.	771	1445.568	Neville
19.	910	1425.021	fred
21.	733	1374.321	mcgonagall
22.	631	1183.078	umbridge
23.	607	1138.080	gryffindor
29.	496	905.889	moody
30.	470	881.215	dobby
31.	502	874.255	Vernon
32.	478	872.288	fudge
34.	433	770.306	Dudley
36.	402	753.720	slughorn
39.	776	681.447	george
41.	410	677.758	percy
44.	353	661.849	slytherin
47.	343	643.100	luna
50.	315	590.602	cedric
51.	312	584.977	kreacher
54.	287	538.104	petunia
55.	298	520.164	riddle
56.	276	517.480	filch
57.	333	515.591	crouch
60.	253	474.356	krum
61.	250	468.732	lockhart
62.	243	455.607	fleur
64.	242	453.732	tonks
65.	235	440.608	goyle
67.	232	434.983	crabbe
68.	231	433.108	bagman
69.	224	419.983	dursleys

75.	212	397.484	cho
76.	204	382.485	Hedwig
77.	199	373.110	bellatrix
79.	195	365.611	wormtail
81.	190	356.236	albus
83.	180	337.487	azkaban
84.	235	337.461	draco
85.	304	337.277	bill
87.	210	333.102	dean
90.	248	326.486	trelawney
91.	180	326.089	seamus
98.	163	305.613	peeves

Table 12: Proper names divided by type

type of proper name	keywords	number of items
first name	Harry, Ron, Hermione, Sirius, Ginny, Neville, Fred, Vernon, Dudley, George, Percy, Luna, Cedric, Petunia, Fleur, Cho, Bellatrix, Albus, Draco, Bill, Dean, Seamus	22
last name	Dumbledore, Hagrid, Snape, Weasley, Malfoy, Potter, Lupin, McGonagall, Umbridge, Moody, Fudge, Slughorn, Riddle, Filch, Crouch, Krum, Lockhart, Tonks, Goyle, Crabbe, Bagman, Dursleys, Trelawney	23
nickname	Voldemort, Wormtail	2
place name	Hogwarts, Azkaban, Gryffindor*, Slytherin*	4
name of magical creature	Dobby, Kreacher, Hedwig, Peeves	4
Total number		55

* also occasionally used as last names

5.3.1 First names

It may not be striking that it is exclusively the positive characters whose first names are key, the only exception and thus a villain that falls into this category is *Bellatrix* who is, nevertheless, sister to one of the important (and positive) characters. Furthermore, the characters who first names are key are only the protagonists, their closest family members and their classmates, who have the same social status and are therefore referred to by their first names. When addressing the students, the Hogwarts teachers occasionally use first names (e.g. Dumbledore addressing Harry) which suggests a closeness of their relationship but more frequently last names are used. Interestingly enough, Dumbledore's first name *Albus* is the only first name of a teacher appearing on the list. This is partly due to his character undergoing dramatic change from the position of unquestioned identity to the status of fallible and imperfect figure. High number of occurrences in the final sequel of the series is also partly due to his full name appearing in the name of a book called *The Life and Lies of Albus Dumbledore.*

5.3.2 Last names

The list of retrieved last names is far more heterogeneous. Quite predictably, it includes the last names of teachers or persons working for the school (*Dumbledore, Hagrid, Snape, Lupin, McGonagall, Umbridge, Moody, Slughorn, Filch, Lockhart, Trelawney*), the formal teacher-student relationship requires the former to address the latter exclusively by their last names, even Harry who has very intimate relationship with Hagrid never addresses him by his first name, Rubeus. The list also includes the last

names of persons of holding a position of authority (e.g. political – *Fudge, Crouch*) and antagonists (*Malfoy, Riddle, Goyle, Crabbe*) or characters that are observed with certain level of suspicion (*Krum, Bagman, Dursleys*). Not surprisingly, Harry's and Ron's family last names appear on the list which is both due to them being numerous in members and very frequent subject of address (the narrative voice almost exclusively uses their first names). Last but not least, a character which hardly falls into any of the previously mentioned subgroups but "prefers to be know by her surname only" /HP5/ (*Tonks*) has a last name with high value of keyness, proving that keyword analysis is in agreement with what is being said in the text.

5.3.3 Other proper names

Since the rest of the subgroups of proper names included only very few members, they will be discussed all together. The nicknames, a product of Rowling's invention, are also quite a predictable item to be included on the keyword list and the same goes for the names of the magical creatures. In order to draw an appealing and convincing picture of an extraordinary a magical world, so different from the reality as we know the author simply cannot use ordinary names. *Gryffindor* and *Slytherin* are both names of the school's colleges but also last names of their founders, their appearance on the keyword list suggests that the membership to one of these houses (*Gryffindor* – noble, *Slytherin* – villainous) is important for the wizarding society since the names of the other houses with equally odd names do not appear this high on the list. Including of two place names (*Hogwarts, Azkaban*) may be yet again linked to the fact that the Series' setting is highly localized and these are products of Rowling's authorial invention, unlike non-key *London, England* etc. that also frequently occur in the novels.

6. Concluding remarks

The objective of this paper was twofold: the first part aimed to provide a brief overview of the recent development in keyword analysis research and most importantly establish the role of the keyword analysis in linguistic exploration of literary works. The second part of this paper attempted to perform the keyword analysis on the Harry Potter series comparing it with other works of children's literature, thus testing the method described in the earlier part, focusing equally on the resultant keywords and exploring various tools of the AntConc program used in the process.

Performing the analysis confirmed some of the previous findings, especially those regarding the processing of the text and the technical constraints of the method. Sardinha (1999) and Culpeper (2009) both warn of the immense load of data retrieved, which often becomes impossible for the researcher to process; an observation which proved true while dealing with the resulting keywords and especially the generated concordance lines and clusters. In this study, the problem was dealt with by sampling – e.g. selecting certain number of concordance lines from each novel, etc.

The keywords retrieved from the corpora suggested several fields in which the Potter series differs significantly from the similarly targeted works of children's literature:

More complex temporal system, very detailed depiction of fictional landscape built on the structure of frequent place relators, direction markers and prepositional clusters, prominence of naturally depicted dialogical speech and focus on the central character's perception and train of thought. Place relators help the reader to visualize the story more vividly, realistic dialogues add credibility to the story and words working to provide the insight into Harry's head aid the audience to put themselves into the character's shoes. Along with the author's choice of relatively common lexis which allows speedy reading, these are all factors that may significantly contribute to the Series' attractiveness and readability.

It is important to note that such keyword analyses are meant to serve as pilot studies, paving the way for further stylistic analyses of the text. Potential incompleteness is therefore their inherent nature, but they bring about new possibilities of quantitative research of literary works, allowing rapid detection of counter-intuitive linguistic patterns which work together with the theme-bound words to create the meaning of a text.

7. **Shrnutí práce v češtině** (Summary of the thesis in Czech)

Práce se zabývá využitím analýzy klíčových slov při stylistickém zkoumání literárních textů. Je rozdělena do dvou hlavních částí: část první pokrývá teoretický základ, který definuje předmět výzkumu, shrnuje dosavadní poznatky v této oblasti a dále obsahuje metodologický oddíl, který se zabývá nastíněním pracovního postupu a popisem jednotlivých funkcí užitého konkordančního programu AntConc. Druhá část představuje vlastní výzkum, při kterém byla generována a následně analyzována klíčová slova ze série sedmi knih o Harry Potterovi, která byla získána za užití referenčního korpusu, vytvořeného z literárních textů pro obdobnou věkovou skupinu.

První část studie definuje a vymezuje pojem klíčových slov (keywords) pro účely této práce. Ta jsou zde chápána jako "slova vyskytující s významně vyšší frekvencí v jednom korpusu v porovnání s korpusem dalším" (Hoffmann et al., 2008: 203). Klíčová slova jsou obecně významná pro význam textu a jsou odrazem toho, čím se daný text skutečně zabývá a jak je strukturně komponován. Dále je často uváděna definice čistě kvantitativní, která je chápe jako určitou statistickou odchylku od jisté normy. Klíčová slova jsou důležitá zejména z hlediska stylistiky, neboť se významně podílejí na stylistické výstavbě a struktuře textu a mohou být dobrým vodítkem pro další analýzu textu. Je také nutné podotknout, že tato práce klíčová slova nechápe a nezabývá se jimi tak, jak je termín užíván v běžném jazyce, sféře kulturních studií, sociologii, knihovnictví a managementu dat a informací apod. – jedná se tedy o pojetí termínu z čistě korpusově-lingvistického hlediska.

Významným faktorem ovlivňujícím výsledná klíčová slova je volba referenčního korpusu. Klíčovost je komparativní kvalita, závislá na kontrastu mezi dvěma skupinami dat, referenční korpus tedy slouží jako filtr pro oddělení slov, která jsou předmětem zájmu od zbytku dat. Ač volba referenčního korpusu do značné míry závisí na cíli

jednotlivých prací, obecně se doporučuje několikanásobně větší objem dat (obvykle bývá referenční korpus alespoň pětkrát větší než korpus zkoumaný). Nejčastěji užívanými statistickými metodami měření „klíčovosti“ jsou potom chi-square nebo log-likelihood test.

Získaná klíčová slova je možno rozdělit do několika podskupin. Opět se zde setkáváme s více přístupy, nicméně základním postupem zůstává dělení do dvou nestejně velkých skupin - na slova nesoucí informační obsah (lexikální) a slova funkčí sloužící k strukturní výstavbě textu (gramatická). Halliday (1973,1978,1994) rovněž rozlišuje “interpersonal keywords”, tedy klíčová slova významná pro organizaci mezilidské komunikace a vyjádření emotivních postojů. Tato práce se nicméně drží dělení základního, ač chápe vlastní jména jako samostatnou podkategorii slov lexikálních. Klíčová slova je též možno dělit na pozitivní a negativní v závislosti na tom, zda jsou statisticky velmi významná nebo ve zkoumaném korpusu naopak téměř chybí.

Analýza literárního textu pomocí získaných klíčových slov má oproti intuitivnímu stylistickému rozboru řadu výhod: díky modernímu softwaru vyvinutému v posledních letech jde o velmi rychlý proces, který je díky nastavitelným parametrům jednotlivých programů možno různě upravovat dle potřeb výzkumu. Komputerizace výzkumného procesu rovněž umožňuje zaujmout kvantitativní přístup k delším úsekům textu, které v minulosti zůstávaly mimo zájem badatelů zejména pro objem dat, která bylo třeba zpracovat. Pravděpodobně nejvýznamnější výhodou oproti klasické stylistické analýze je nicméně možnost identifikace pro text charakteristických vysokofrekvenčních gramatických slov, která tvoří nosnou strukturu textu a je téměř nemožné je odhalit pomocí intuice. V neposlední řadě je této metodě připisována schopnost do jisté míry zredukovat subjektivitu a dodat jazykové analýze systematičnost. Výsledná data mohou sloužit jako pilotní studie pro hlubší stylistickou analýzu, mohou být využita v překladu i při odhalování autorství literárních děl.

Za nevýhody této metody se naopak považuje její novost a nedostatek poznatků v oblasti, částečně daný neustálým vývojem nových technologií. I když analýza klíčových slov subjektivitu při výzkumu významně snižuje (představuje metodu typu „corpus-driven“), nelze ji provést bez přítomnosti lidské práce a finální produkt je vždy osobní interpretací konkrétního výzkumníka. Identifkace klíčových slov je prvním krokem k detailnější analýze kolokací, frazeologie nebo sémantické prozodie. Dalším problémem je obrovské množství vygenerovaných dat, jejichž zpracování je často nad možnosti dané práce a také to, že výsledná klíčová slova poukazují pouze na rozdíly v lexiku, nikoliv na shodné prvky.

Závěrečný oddíl teoretické části se věnuje dosavadním studiím v oblasti výzkumu. Lingvistické analýzy poetických textů jsou i z historického hlediska poměrně častým jevem, v případě textů prozaických jde o postup využívaný až v posledních letech. Analýza klíčových slov je jedním ze základních postupů korpusové stylistiky a je založená na myšlence, že myšlenka a estetická hodnota textu je v praxi neoddělitelná od formy, tedy jazykových struktur, ve kterých je zapsána. V minulosti využili tento postup

například Burrows (1987), který zkoumá jazyk postav v románech Jane Austen, Tribble (2000), který se zaměřuje na romantickou prózu a Culpeper (2009), který analyzuje jazyk postav v Shakespearovské tragédiii. Na sérii knih o Harry Potterovi se soustředí studie Čermákové a Fárové (2010), ale na toto téma existuje i řada lingvistických studií které volí jiné postupy než využití klíčových slov.

Následující metodologická část shrnuje a odůvodňuje zvolenou metodu. Jak už bylo řečeno, základním výzkumným postupem pro tuto práci bylo generování klíčových slov ze série knih o Harry Potterovi za užití referenčního korpusu skládajícího se z textů extrahovaných z Britského národního korpusu, pocházejících z přibližně stejné doby a určených pro stejnou věkovou skupinu. Vzhledem k rozsahu práce byl počet analyzovaných klíčových slov stanoven na sto. Pro analýzu byl použit konkordanční software vyvinutý Laurencem Anthonym, který kromě vyhledání klíčových slov (Keyword List tool) obsahuje i několik dalších funkcí, které byly při výzkumu užity: jde zejména o funkce generování jednoduchého frekvenčního seznamu slov (Word List tool), zobrazení konkorančních řádků (Concordance tool), vyhledání širšího kontextu (Concordance Plot tool), detekce shluků lexikálních jednotek (N-Grams) a detekce shluků lexikálních jednotek v okolí konkrétního klíčového slova (Clusters tool).

Úvod praktické části se zabývá dělením získaných klíčových slov, které se neobešlo bez komplikací, zejména kvůli existenci řady anglických slov, z jejichž morfologické struktury se nedá určit slovnědruhová příslušnost. Následující oddíl se zabývá analýzou výsledných gramatických klíčových slov. Vzhledem k tomu, že hlavní hrdina a většina dalších důležitých postav jsou muži, významnou podskupinu představují mužská zájmena osobní a přivlastňovací. Jakkoli je vysoká míra užití v podobném textu předvídatelná, generování lexikálních shluků kolem zájmena *he* umožnilo odhalit, že vysoká prominence zájmena je mimo jiné dána jeho zahrnutím do přezdívky antagonisty *he-who-must-not-be-named*. Neobvykle velká frekvence tázacího a vztažného zájmena *who* je z velké části rovněž opodstatněna touto přezdívkou. Dalším zajímavým zjištěním je časté spojení zájmena *he* se slovy označujícími mentální procesy či lidské vnímání. Důležitou skupinu mezi gramatickými slovy představují předložky. Objevují se zde ne zcela obvyklá gramatická uskupení – často se vyskytují sekundární předložky, obvykle v kombinaci s adverbii nebo dalšími předložkami a substantivy označujícími místa v prostoru, což reflektuje velmi konkrétní zasazení příběhu ve fiktivní krajině a autorčin cit pro jemné rozdíly v poloze nebo směru a pro celkové prostorové rozmístění postav a předmětů. Zajímavé, leč co do výskytu méně významné prvky představují částice vyjadřující souhlas (*okay, yeah*) a citoslovce váhání (*er*), které potvrzují častý výskyt živého realistického dialogu a autorčinu schopnost věrně napodobit prozodii mluveného slova.

Lexikální klíčová slova byla rovněž klasifikována dle sémantického obsahu. Nejpředvídatelnější sémantickou doménou jsou slova označující názvy osob. Z nich některá odrážejí zasazení příběhu do akademického prostředí (*students, professor, madam*), kde existuje systém formálního oslovování mezi vyučujícími a studenty, další jsou slova se zobecněnou referencí, odkazující na základní kategorie (*wizard, wizards*).

Zajímavé je, že tato slova můžeme chápat v kontextu příběhu Harryho Pottera jako obecné termíny mající přibližně stejnou referenční funkci, jako má v reálném světě *person, people*. Podobně předvídatelná jsou slova, která se svým obsahem vztahují ke kouzelnictví a magickým předmětům. O mnoho zajímavější jsou podstatná jména označující fyzická místa (*hall, grounds, office, ministry*), která tvoří lexikální komplement frekventovaným předložkám, a opět tak potvrzují hypotézu o velmi konkrétním zasazení příběhu v prostoru. Prostorové relátory ou spolu s dalším kohezními prostředky určující pro orientaci čtenáře v textu. Tato sekce dále zahrnuje podstatná jména magických bytostí (*Death Eaters, Dementors, Muggles*), která jsou zde zahrnuta zejména proto, že jsou často užita s neurčitým členem *a*, a významově spíše zařazují jedince do konkrétní třídy. Jediné klíčové adverbium *slightly* na základě získaných kolokací v sérii často modifikuje sloveso v popisu souběžných akcí a slouží k upřesnění směru či rozsahu pohybu (*moving slightly to his right*). Vzhledem k tomu, že je text vyprávěný v minulém čase a třetí osobě, je sloveso *said* předvídatelně frekventovaným slovem, nicméně v porovnání s podobně strukturovanými texty velmi překvapivě slovem klíčovým – jeho klíčovost opět potvrzuje častou přítomnost dialogu. Jako poslední je nutno jmenovat sloveso *looking*, které kromě svého základního významu „hledět“ v textu často nabývá v nefinitní formě funkci označení vedlejšího děje či sponového slovesa typu *seem*, které může být chápáno jako další ukazatel (vedle *he*) k tendenci příběhu zaměřovat se na smyslové vnímání protagonisty(ů).

Jako s podskupinou lexikálních slov bylo v rámci této studie zacházeno s vlastními jmény. Ta jsou uskupením předvídatelným, neboť jsou víceméně identická se jmény protagonistů. Poněkud zajímavější jsou však jejich formy: klíčová křestní jména jsou až na jednu výjimku výlučně jmény postav kladných a jde pouze o jména protagonistů, členů jejich rodin a spolužáků – členství ve stejné společenské skupině dovoluje jednotlivým postavám vzájemně se oslovovat křestními jmény. Příjmení označují osoby různorodé – jde zejména o vyučující (je zde implikován formální vztah), postavy v pozici autority (ministr) a antagonisty, popř. postavy vnímané spíše negativně. Další slova spadající do této podskupiny představují přezdívky, místní jména, jména zvířat a bytostí, popř. jména institucí.

V závěru práce jsou shrnuty užité postupy a materiály, jejich efektivnost a konkrétní problémy, se kterými se bylo nutno vypořádat v průběhu vypracování práce a popřípadě jejich řešení. Je v ní rovněž obsažen stručný komentář ohledně užití a výhod metody, které se při jejím testování ukázaly jako významné a její užití pro stylistickou analýzu opodstatňují, a taktéž úskalí, kvůli kterým některé její prvky pro stejnou oblast zkoumání nelze bez výhrad doporučit. Poslední složkou je shrnutí výsledků praktické části, které poukazují na prvky, kterými si série knih o Harrym Potterovi od ostatní literatury výrazněji liší (komplexnjší využití časových rovin, detailní popis fiktivní krajiny a cit pro směr a orientaci v prostoru, velmi frekventovaný a realistický dialog, soustředění se na smyslové vnímání a mentální pochody protagonistů atd.) a mohou tak významně přispívat k jeho čtivosti a popularitě.

List of references and sources

References

Anthony, Laurence. *AntConc (Windows, Macintosh OS X, and Linux) Build 3.2.2.1.* Waseda University. Web. Last accessed on August 13, 2011. < http://www.antlab.sci.waseda.ac.jp/software/README_AntConc3.2.2.1.pdf>

Archer, D. (2009) *What's in a word-list?: investigating word frequency and keyword extraction*. Farnham: Ashgate Publishing.

Baker, P. 2004. Querying keywords. Questions of difference, frequency and sense in keyword analysis. *Journal of English Linguistics* 32(4): 346–359.

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London and New York: Continuum.

Berber Sardinha, T. 1999. Using key words in text analysis: Practical aspects. *DIRECT Papers* 42, LAEL, Catholic University of São Paulo.

Brett, David Finbar (2009) *Eye dialect: translating the untranslatable.* Annali della Facoltà di Lingue e Letterature Straniere di Sassari, Vol. 6 , p. 49-62.

Bondi, M. &M.Scott (eds) (2010) *Keyness in texts*. Amsterdam&Philadelphia: John Benjamins.

Čermáková, A. & Fárová, L. (2010). *Keywords in Harry Potter and their Czech and Finnish Translation Equivalents*. In: F. Čermák, A. Klégr, P. Corness (eds.): "InterCorp: Exploring a Multilingual Corpus". Praha: NLN / Ústav Českého národního korpusu. Pp 177-188.

Culpeper, J. (2009) 'Keyness. Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*', *International Journal of Corpus Linguistics* 14:1, pp 29-59.

Diesing, M. (2007) *How To Do Things with Words and Wands: The pragmatics of casting spells*. Cornell University. Web. Last accessed on August 13, 2011. <http://dingo.sbs.arizona.edu/~ hharley/PDFs/Blog/DiesingWandW.pdf.>

DuškováL.; et al. (1994) *Mluvnice současné angličtiny na pozadí češtiny*. Praha: Academia.

Firth, J. R. 1935. Technique of semantics. *Transactions of the Philological Society*, 36–72.
Fischer-Starcke, B. (2009) 'Keywords and frequent phrases of Jane Austen's Pride and Prejudice', *International Journal of Corpus Linguistics* 14:4, pp 492-523.

Halliday, M. A. K. (1973) *Explorations in the Functions of Language*. London. Edward Arnold.

Halliday, M. A. K. & Hasan, R. (1976) *Cohesion in English*. London: Longman.

Halliday, M. A. K. (1978) *Language as Social Semiotic: The Social Interpretation of Language and Meaning.* London: Edward Arnold.

Halliday, M. A. K. (1994) *An Introduction to Functional Grammar*. (2nd edition). London: Edward Arnold.

Hoffmann, S. et al. (2008) *Corpus Linguistics with BNCweb – A Practical Guide*, Frankfurt am Main: Peter Lang.

Mahlberg, M. (2007) Clusters, key clusters and local textual functions in Dickens. *Corpora*, 2(1), 1-31

Mahlberg, M. (2009) *Lexical cohesion and corpus linguistics*. Amsterdam&Philadelphia: John Benjamins.

Leech, G. and M. Short (2007, 1st ed. 1981) *Style in Fiction*, London: Longman.

McEnery, T. and A. Wilson (2001, 1st ed. 1996) *Corpus Linguistics. An Introduction*. Edinburgh: Edinburgh University Press.

Nygren, A. (2006) *Essay on the Linguistic Features in J.K.Rowling's Harry Potter and the Philosopher's Stone*. Stockholm University. Web. Last accessed on August 13, 2011. <http://www.essays.se/essay/b2de0726c7/>

Quirk, R; et al. (1989) *A Comprehensive Grammar of English Language*. New York: Longman.

Rayson, P. (2008) 'From key words to key semantic domains', *International Journal of Corpus Linguistics* 13:4, pp 519-549.

Scott, M. R. (2000) *Focusing on the text and its key word*s. In L. Burnard & T. McEnery (Eds.), "Rethinking Language Pedagogy from a Corpus Perspective", Volume 2. Frankfurt: Peter Lang 103-122.

Scott, M. R. & Tribble C. (2006) *Key Words and Corpus Analysis in Language Education*. Amsterdam&Philadelphia: John Benjamins.

Scott, M.R. (2008) *WordSmith Tools Help Manual*. Version 5.0. Liverpool: Lexical Analysis Software.

Scott, M.R. (2010) *Problems in investigating keyness, or clearing the undergrowth and marking out trails*. In M. Bondi and M.R. Scott (Eds.) "Keyness in texts". Amsterdam&Philadelphia: John Benjamins.

Tognini Bonelli, E. (2001) *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Sinclair, J. McH. & Mauranen, A. 2006. *Linear Unit Grammar* [Studies in Corpus Linguistics]. Amsterdam: John Benjamins.

Stubbs, M. (2005) 'Conrad in the computer: examples of quantitative stylistics methods', *Language and Literature* 14 (1), pp 5-24.) 10

Teubert W. and A.Čermáková (2007) *Corpus Linguistics: A Short Introduction*. London: Continuum.

Sources

AntConc 3.2.2.1 by Laurence Anthony, Waseda University < http://www.antlab.sci.waseda.ac.jp/software.html>

British National Corpus (BNC), last accessed on August 13, 2011. <http://www.natcorp.ox.ac.uk/>

Oxford Advanced Learner's Dictionary, last accessed on August 13, 2011. <http://www.oxfordadvancedlearnersdictionary.com/>

Rowling, J.K. *Harry Potter and the Sorcerer's Stone*. New York: A.A. Levine Books, 1998.

Rowling, J.K. *Harry Potter and the Chamber of Secrets*. New York: A.A. Levine Books, 1999.

Rowling, J.K. *Harry Potter and the Prisoner of Azkaban*. New York: A.A. Levine Books, 1999.

Rowling, J.K. *Harry Potter and the Goblet of Fire*. New York: A.A. Levine Books, 2000.

Rowling, J.K. *Harry Potter and the Order of the Phoenix*. New York: A.A. Levine Books, 2003.

Rowling, J.K. *Harry Potter and the Half-Blood Prince*. New York: A.A. Levine Books, 2005.

Rowling, J.K. *Harry Potter and the Deathly Hallows*. New York: A.A. Levine Books, 2007.